Technology

How the Internet Archive is waging war on misinformation

San Francisco-based non-profit is archiving billions of web pages in a bid to preserve web history

Camilla Hodgson in San Francisco YESTERDAY

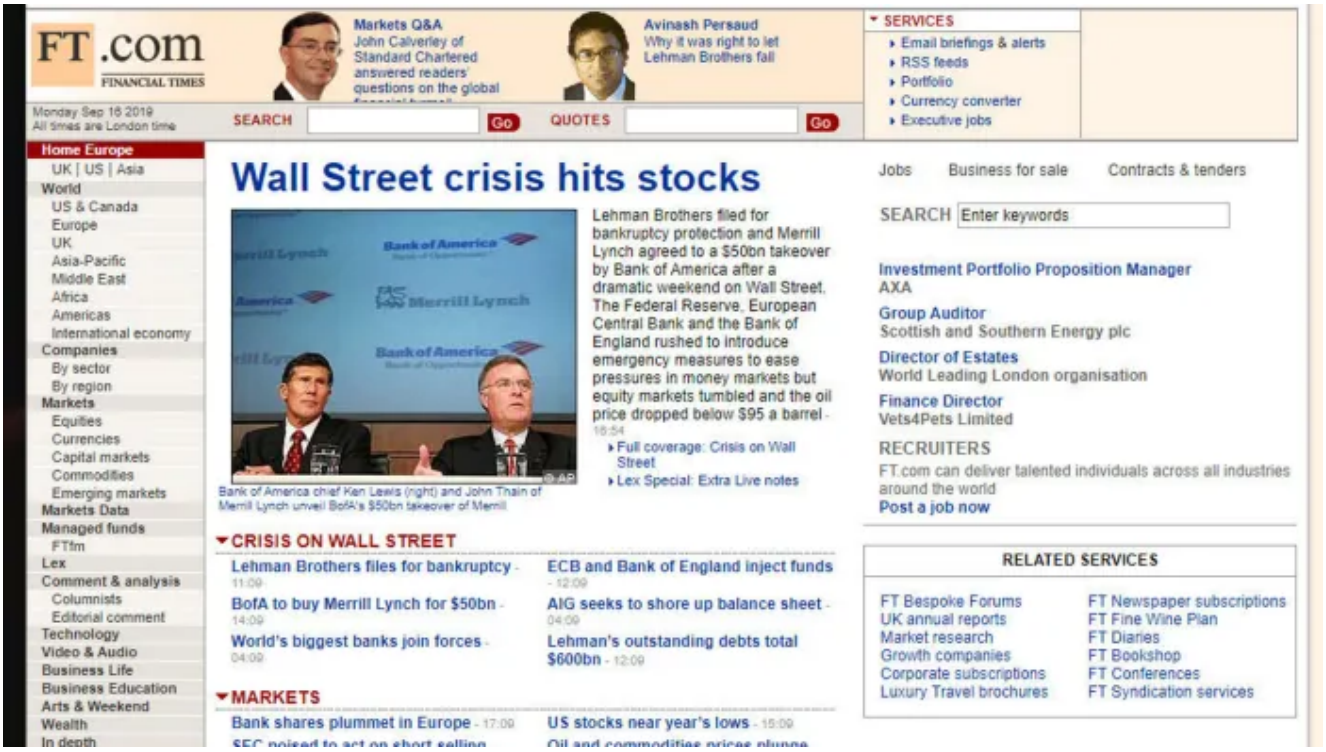
On a foggy September lunchtime in San Francisco, a group of researchers and data scientists sat around foldable plastic tables in what was once a Christian Science church, evangelising about open-source information and the democratisation of knowledge.

The 50-strong party, which had been assembled for a weekly progress discussion, dined downstairs in a pillared building that now houses the Internet Archive, a digital library dedicated to providing "universal access to all knowledge". As computers hummed on cluttered workstations all around, the employees and a handful of invited guests greeted each update with optimistic applause.

The Internet Archive, founded in 1996, is a non-profit that collects and digitises information, from films to books. It is best known for the Wayback Machine, a free repository of web pages that allows users to see what a particular URL looked like when it was archived, regardless of whether it has since been changed or taken down.

Since the 2016 US election, as fears about the power of fake news have intensified, the archive has stepped up its efforts to combat misinformation. At a time when false and ultra-partisan content is rapidly created and spread, and social media pages are constantly updated, the importance of having an unalterable record of who said what, when has been magnified.

"We're trying to put in a layer of accountability," said founder Brewster Kahle.



A screenshot of the homepage of FT.com from September 15 2008, the day Lehman Bros filed for bankruptcy, taken from the Wayback Machine archive © Internet Archive

Mr Kahle founded the archive, which now employs more than 100 staff and costs $18m a year to run, because he feared that what was appearing on the internet was not being saved and catalogued in the same way as newspapers and books. The organisation is funded through donations, grants and the fees it charges third parties that request specific digitisation services.

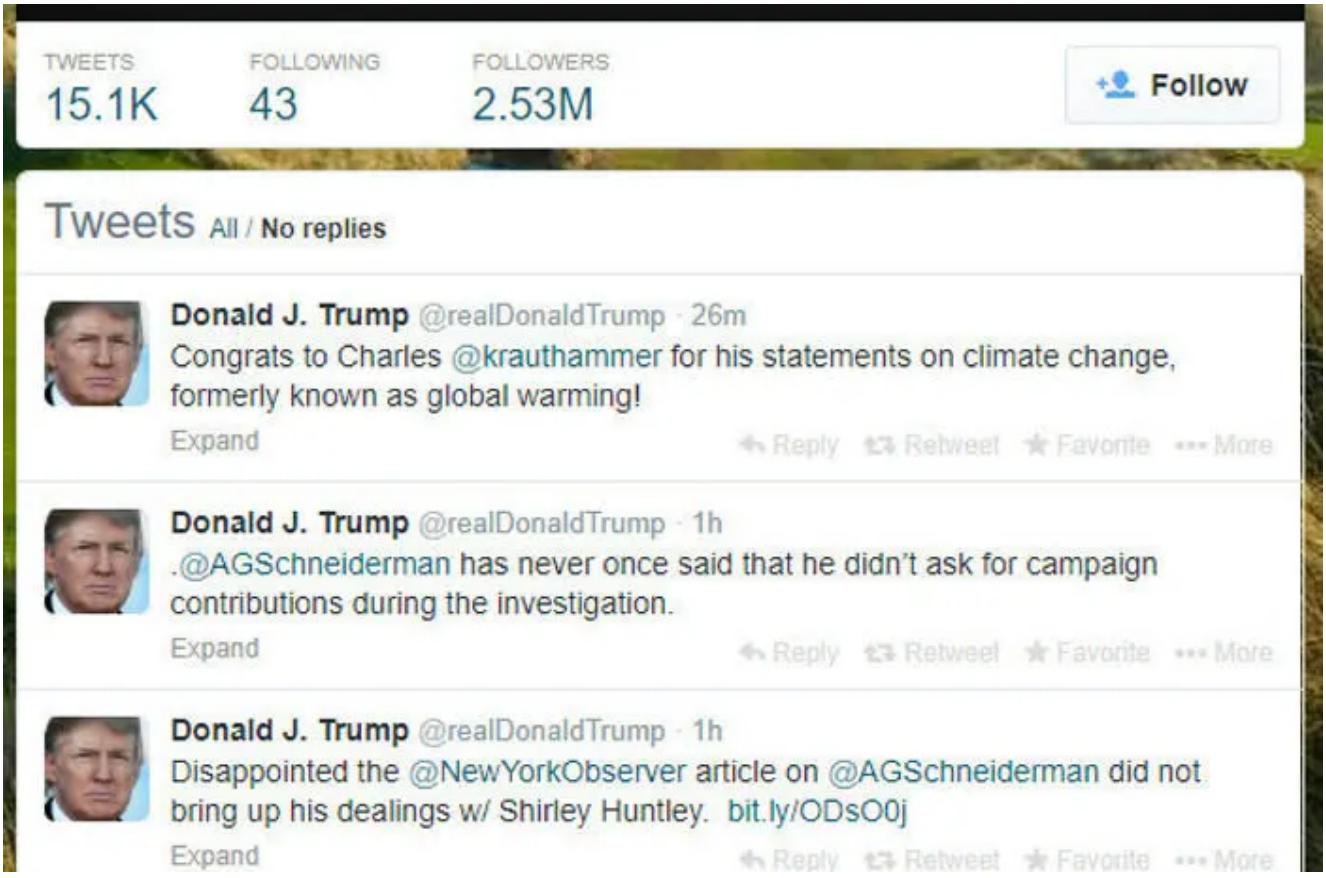
So far, the archive has catalogued 330bn web pages, 20m books and texts, 8.5m audio and video recordings, 3m images and 200,000 software programs. The most popular, public websites are prioritised, as are those that are commonly linked to. Some information is free to access, some is loaned out (if copyright laws apply) and some is only available to researchers.

Curled up in a chair in his office after lunch, Mr Kahle lamented the combined impact of misinformation and how difficult it can be for ordinary people to access reliable sources of facts.

“We’re bringing up a generation that turns to their screens, without a library of information accessible via screens,” said Mr Kahle. Some have taken advantage of this “new information system”, he argued — and the result is “Trump and Brexit”.

Having a free online library is crucial, said Mr Kahle, since “[the public is] just learning from whatever ... is easily available”

After President Trump's election, and as the existence of disinformation campaigns that sought to sway voters came to light, the archive began several new projects. One of these was the Trump Archive, a collection of the US president's television appearances that now contains more than 6,000 videos, including from before he took office. Separately, as part of the effort to document the 45th president's often-contradictory statements, the organisation is cataloguing Mr Trump's tweets.



Tweets by Donald Trump from 2014, as captured by the Wayback Machine © Internet Archive

Social media is "critically important, it's the communication platform of our time", said Mark Graham, director of the Wayback Machine. News feeds on platforms such as Facebook, as well as chat apps, are the "dominant way" many people get information, he said. "That's how they're learning about the world" and "who they think their enemy is".

The archive hopes its repository will help others identify false information and fact-check suspicious content. The emergence of deepfakes — videos that appear to show someone doing or saying something they did not do or say — is a "monster problem", said Roger Macdonald, director of the organisation's TV archive. But having a library of videos means experts and algorithms can help spot those that have been tampered with or taken out of context.

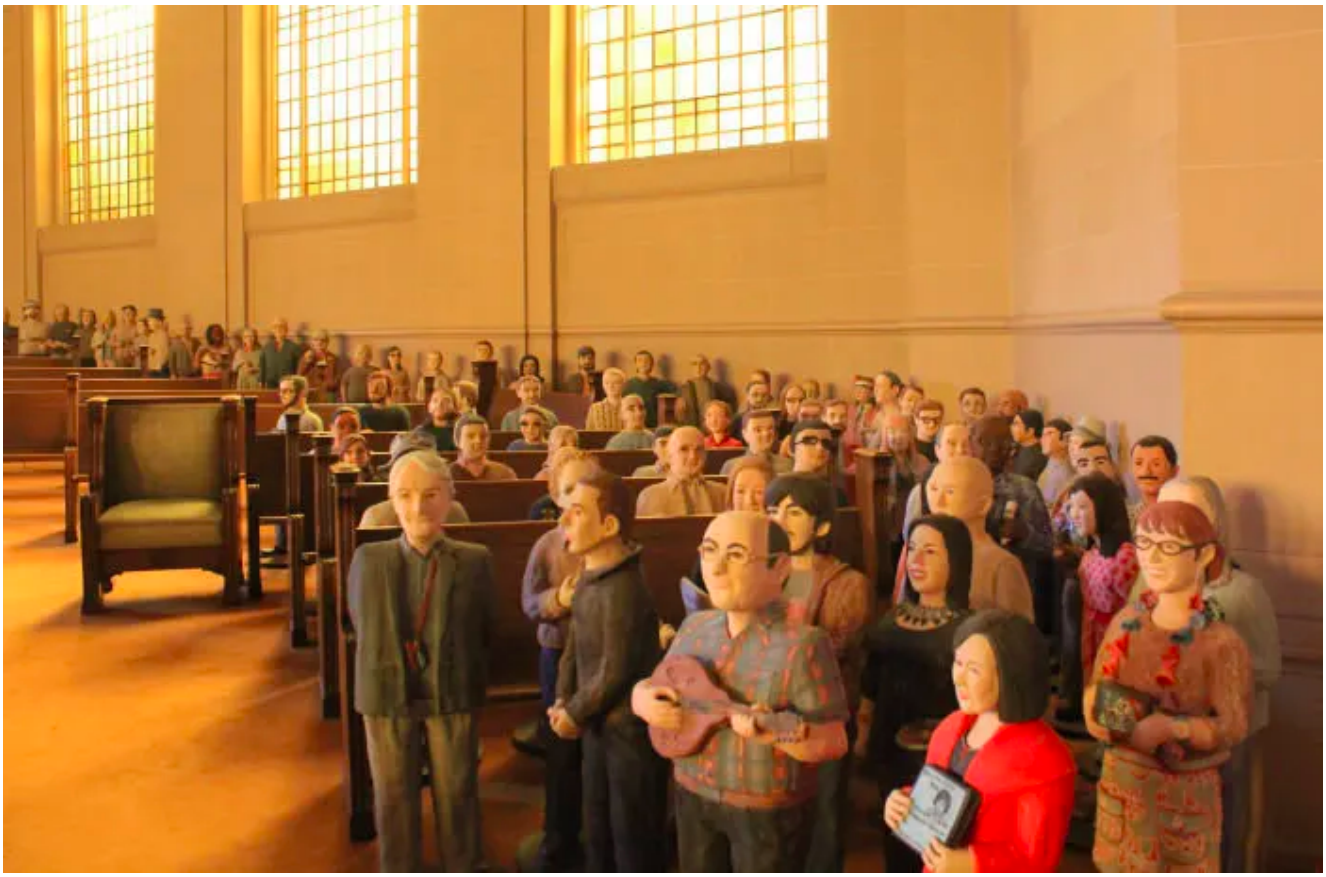
Deciding what to do about fakes is more difficult, and not part of the archive's mandate. But

necessarily the answer. Hateful material need not remain publicly available, he said, but certain researchers and politicians should be able to study it.

As such, the Wayback Machine does not filter out misinformation. “It’s not about trying to archive the stuff that’s true, but archive the conversation. All of that is what people are experiencing,” said Mr Graham.

Given the internet’s explosive growth in the past two decades — there are currently more than 60tn web pages — the task of archiving it has become increasingly difficult. But Mr Kahle said he is hopeful his organisation is keeping up, at least with cataloguing the most popular, public websites.

Mr Graham said he was an “optimist”, but conceded that the archive had not yet saved as much as he would like. Take YouTube, for example: the team is only archiving a “small fraction” of all the videos published each week.



Porcelain replicas of every employee who has spent at least three years working at the Internet Archive, which stand around the edges of the organisation's headquarters © Camilla Hodgson

The organisation currently uses about 3,000 different “crawlers”, algorithms that take regular snapshots of certain public, paywall-free web pages that are stored in the Wayback Machine. Some are very specific, such as political websites from specific geographies, and

some are much broader. Organisations can pay the archive to set up a specific crawler, which around 650 have done.

The archive itself, this partial copy of the internet, is stored in six 6ft-high servers that sit upstairs in what used to be the nave of the church. There is a full back-up copy elsewhere in California, and partial copies in Canada, the Netherlands and Alexandria, Egypt. This is precautionary, said Mr Kahle: remember the burning of the great Library of Alexandria.

Around the sides of the room stand around 130 3ft porcelain figurines — replicas of every employee who has spent at least three years at the archive.

Ultimately, said Mr Macdonald, the Internet Archive is only a starting point. It's "a series of beta tests for what could be done at scale if society really got into it".

Despite its San Francisco roots, Mr Kahle said the non-profit has little in common with modern-day Silicon Valley, where the wealth gap is enormous and a few executives control platforms used by billions. He hopes the "legacy of all this technology" is not that "we have fewer winners," he said. "I like it when lots of people win."

Copyright The Financial Times Limited 2019.
All rights reserved.